
46
FIAT
YAMAHA
FASTWEB
agv
motogp.com
official website

Bringing Astronomers & Computer Scientists Together:
New Methods for Calculating Galaxy Photometric Redshifts in the Sloan Digital Sky Survey

Michael Way (NASA/Goddard Institute for Space Studies)
Paul Gazis, Jeffrey Scargle (NASA/Ames, Space Sciences Division)
Ashok Srivastava (NASA/Ames Intelligent Systems Division)
Les Foster + Students (San Jose State University)
Rama Nemani (NASA/Ames Earth Science Division)

Outline

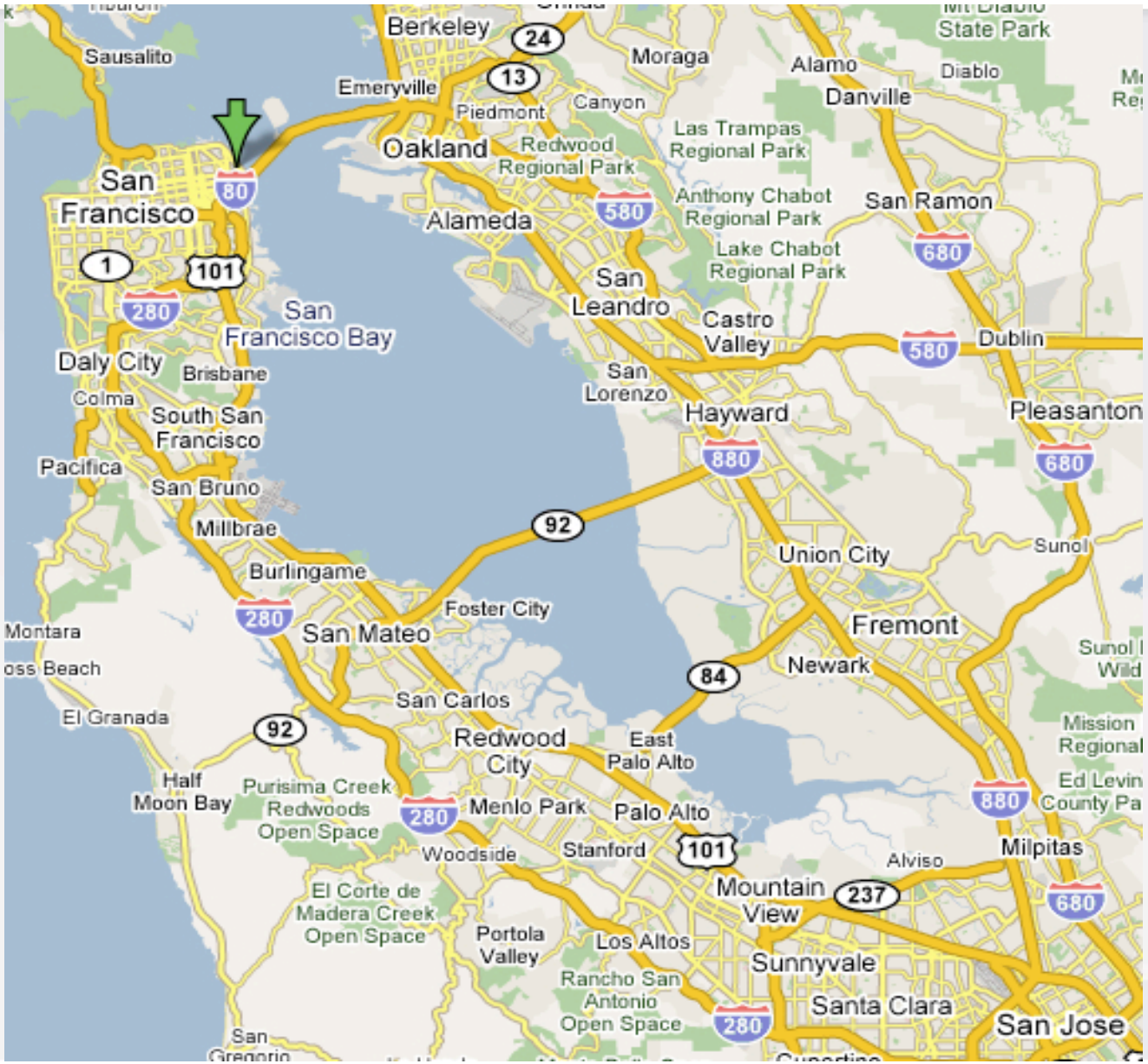
- Astro-CS Collaborations
- Geography
- What good are galaxy redshifts
- Why are spectra “expensive” & How to get them
- What are Photometric Redshifts & How to get them
- The Sloan Digital Sky Survey: Description
- Number Density in the Sloan: Photometry vs Spec
- Photometric Estimation Methods
- Linear Regression & Non-Linear Regression
- Gaussian Process Regression
- Results, The Future, Conclusions

Collaborations Everywhere?

A few well known Astro-CS collaborations:

- **I**nternational **V**irtual **O**bservatory **A**lliance (IVOA)
- **A**stronomical **D**ata **A**nalysis **S**oftware & **S**ystems
- SDSS + Microsoft Research
  - Casjobs: Alex Szalay (JHU) & Jim Gray (MS)
- LSST = Google + Bill Gates + NSF + …
- Penn State Center for Astrostatistics(Summer School)
- Institute for Pure and Applied Mathematics [UCLA]
- (Ames ROSES workshops, Google+Ames seminars, & **many** others…)

The Necessities of Geography?

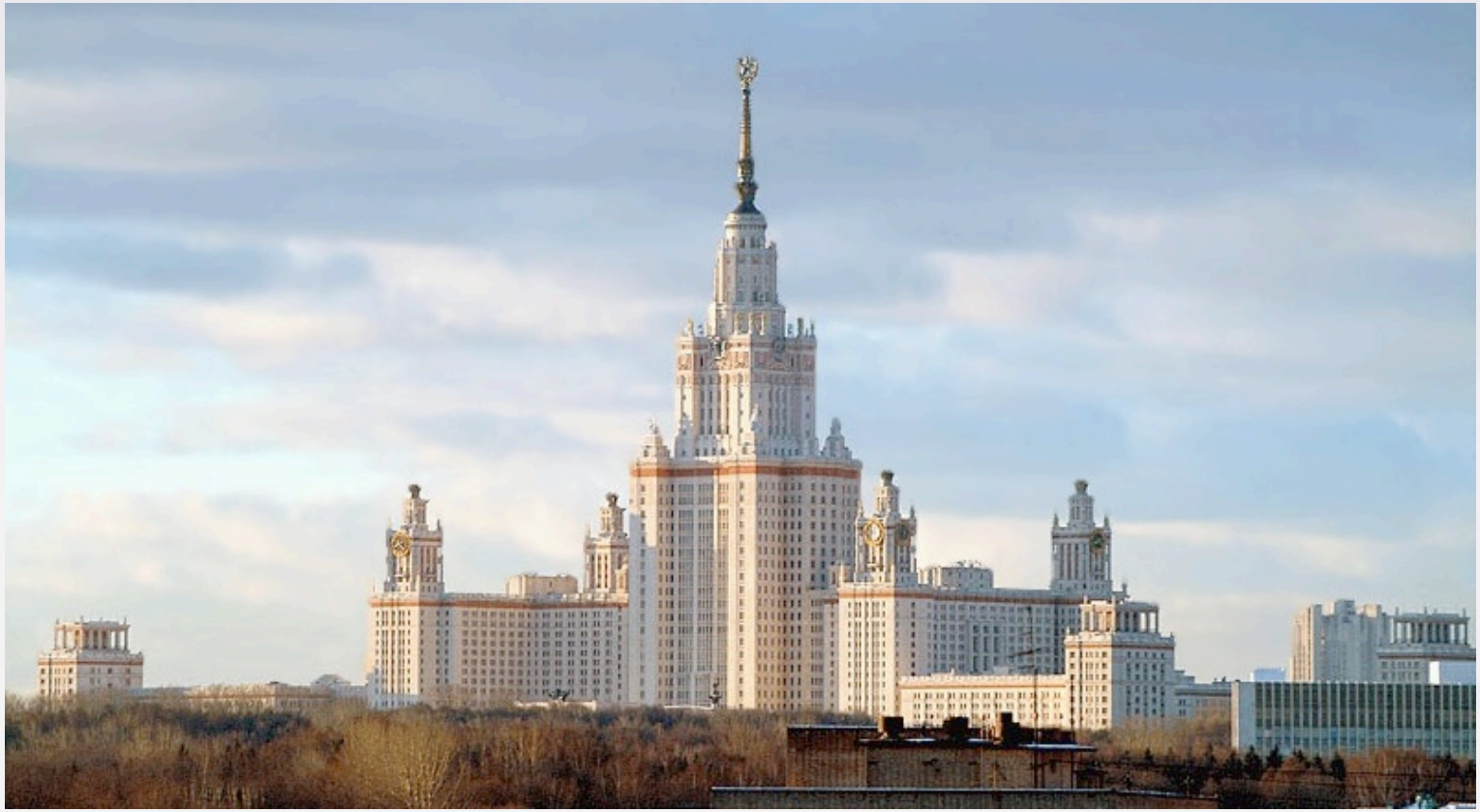


Silicon Valley

Berkeley
Stanford/SLAC
NASA
UC-Santa Cruz/Lick
San Jose State
San Francisco State
Google
Xerox Park
Etc…

The Non-Necessities of Geography

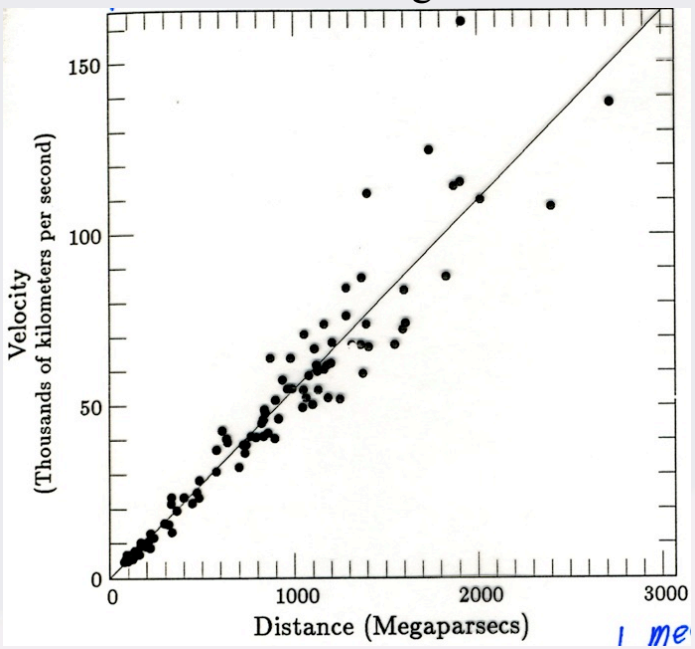
Moscow State University - USSR

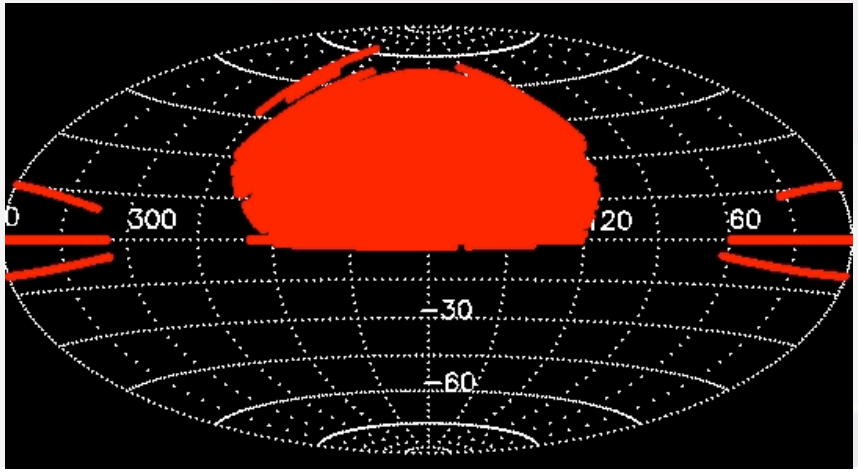


Who Needs Redshifts?

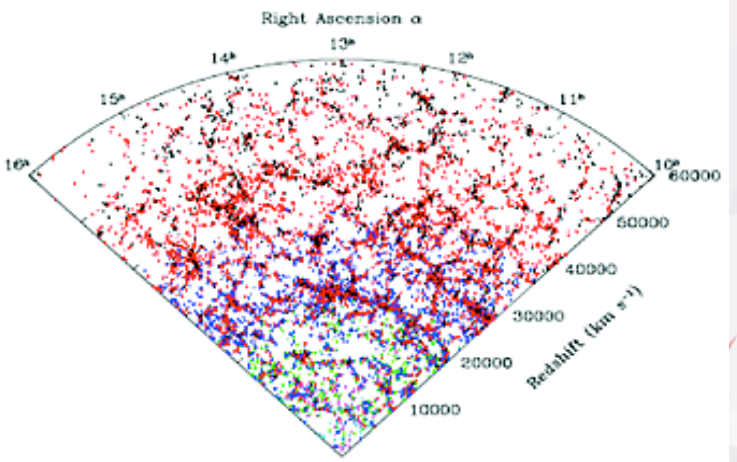
- Since Hubble we have used redshifts as a proxy for distance in the Universe: distance=v_r/H
- They also allow one to constrain formation scenarios for Large Scale Structure in 3-D Cosmological Models

Hubble Diagram





2-D to 3-D Redshifts ➔➔➔



Those expensive spectra

BUT…

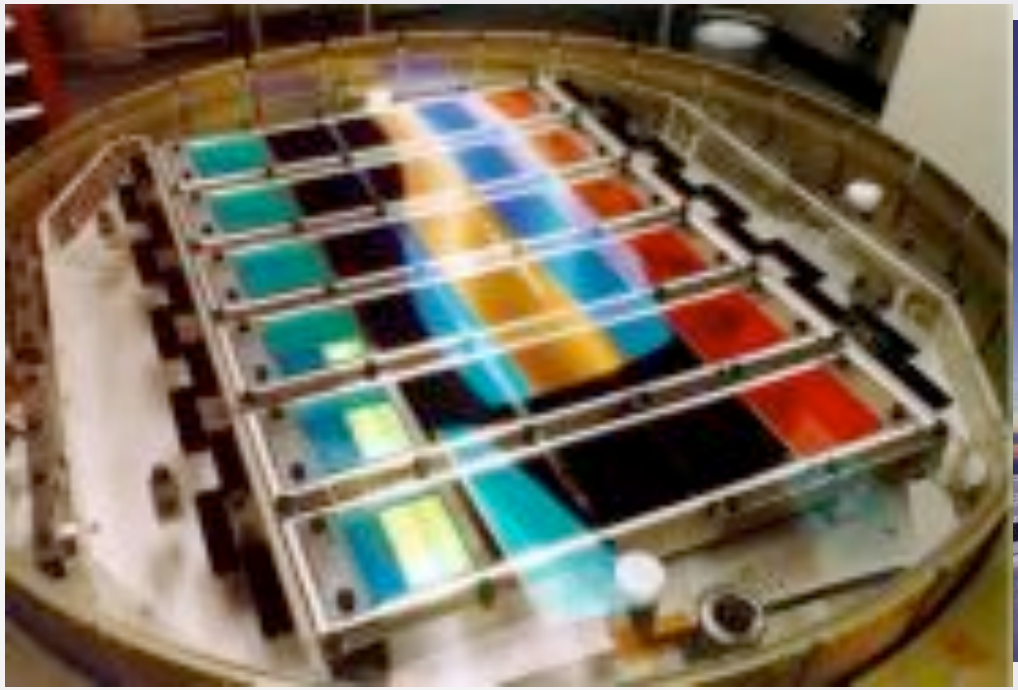
- Spectra for redshifts are expensive to obtain
  For Example: The first CFA catalog?

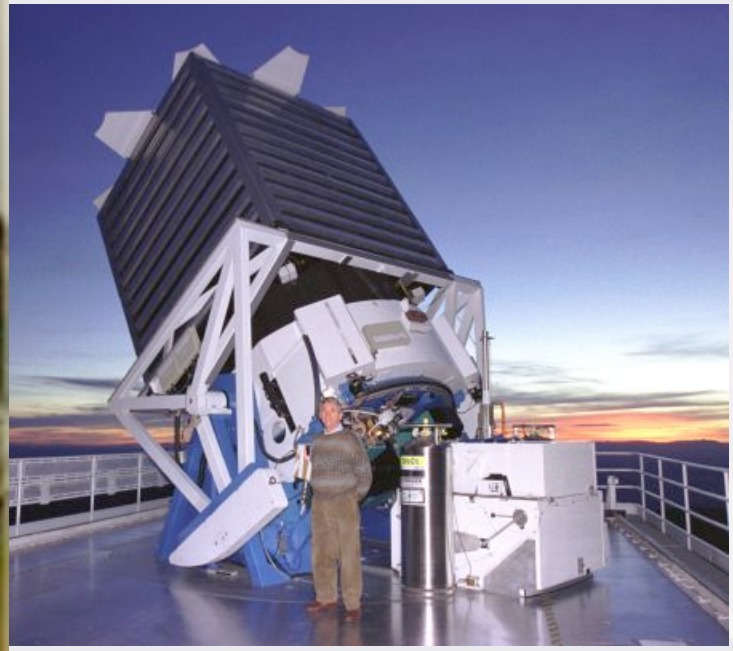
  2401 spectra from the merged Zwicky-Nilson catalog took 5 years to obtain: 1977 - 1982
- Even now measuring a spectrum of sufficient S/N for redshift measurements requires more time than equivalent quality photometry

Practical Considerations: LSS via Spec-z?

Spectroscopic photons are costly (time/resources):

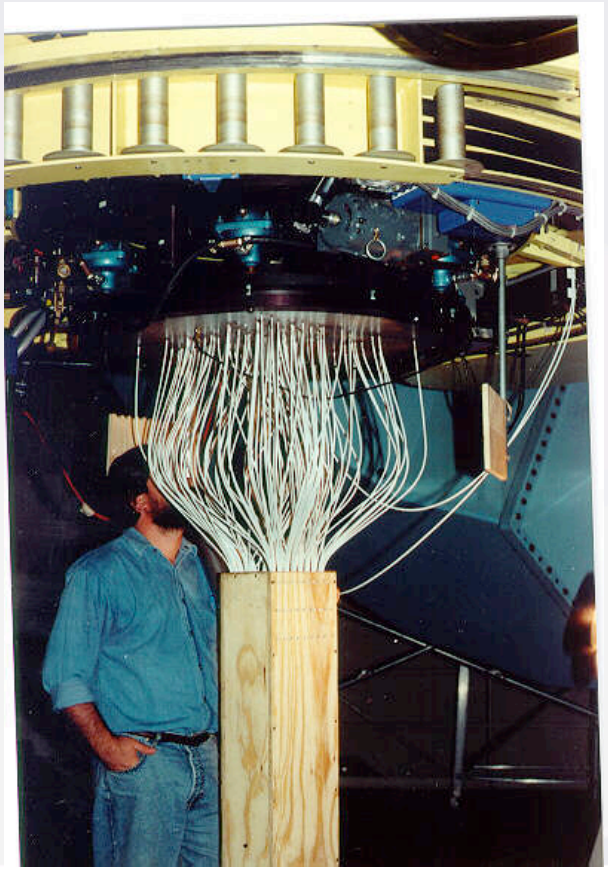
First: The galaxy must be found in an imaging survey





LSS via Spec-z?

Second: Spectra must be measured with a costly specialized instrument



Those expensive spectra

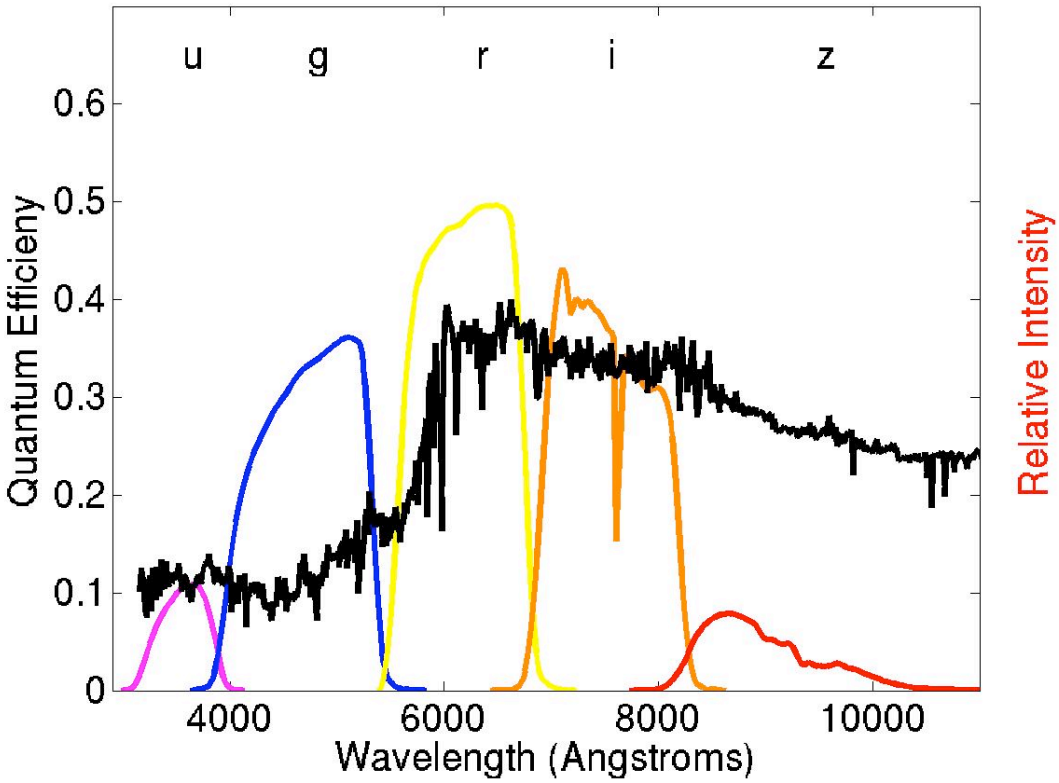
So…

- It would be desirable (if possible) to obtain the redshifts from multi-band photometry alone
- Baum (1962): was the first to attempt "**Photometric Redshifts**" using 9 broad bands
- Lets take a look at Photometric Redshifts

What are Photometric Redshifts?

Photometric Redshifts: A **rough** estimate of the redshift of a galaxy without having to measure a spectrum.

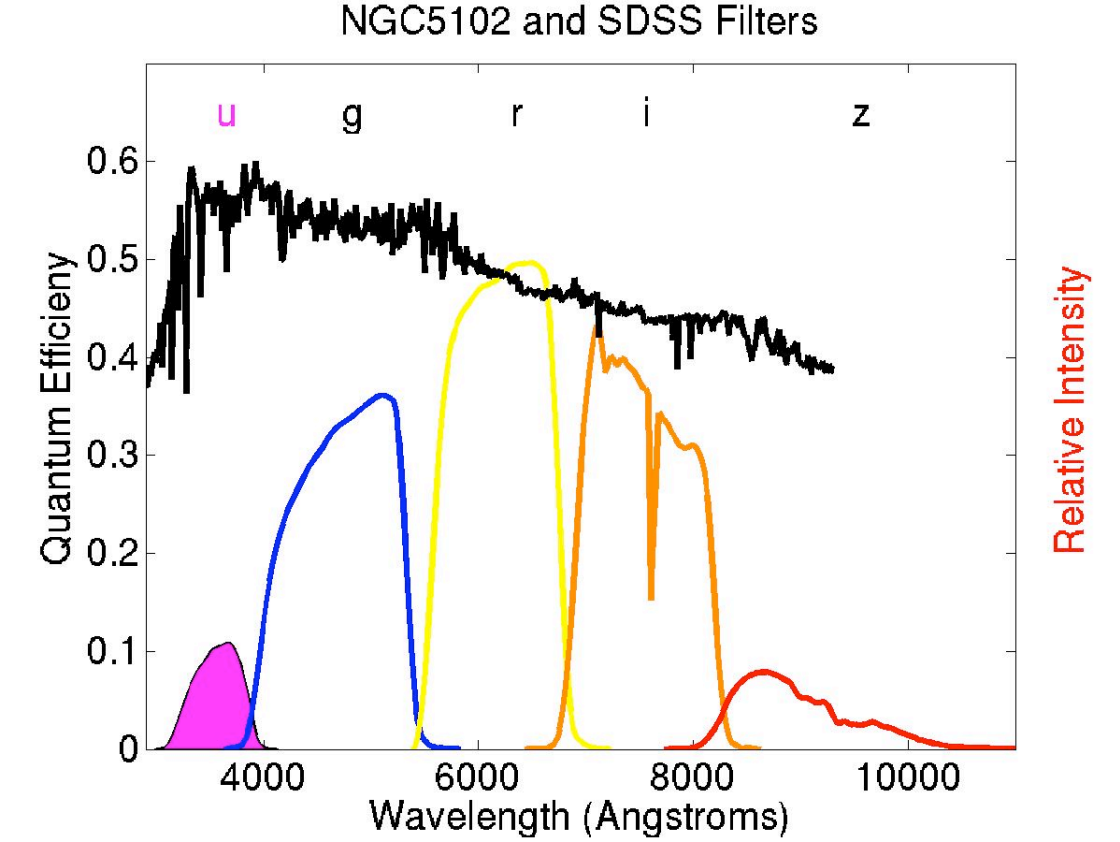
$\mathbf{Z_{spec}=(\lambda_{measured}-\lambda_{rest})/\lambda_{rest}}$ $\qquad$ $z_{photo}=z(C,m)$



What are Photometric Redshifts?

$z_{spec}=(\lambda_{measured}-\lambda_{rest})/\lambda_{rest}$ $z_{photo}=z(C,m)$

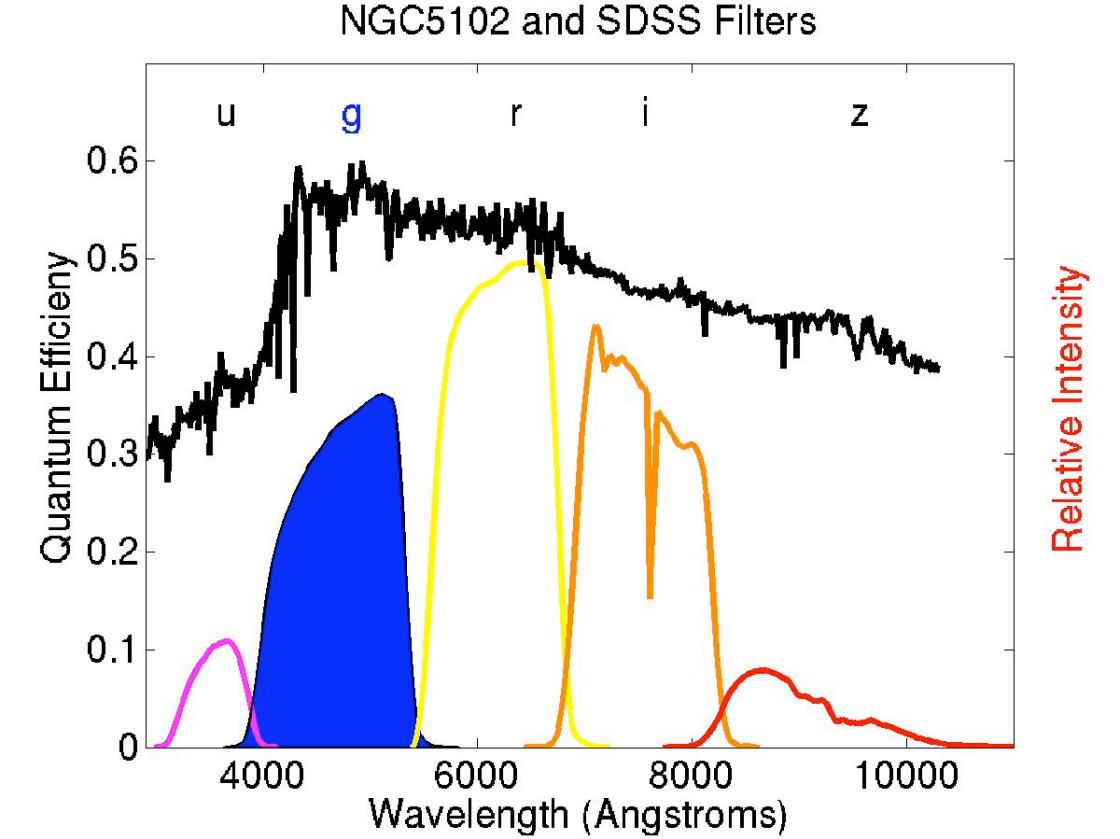
z=0.0



What are Photometric Redshifts?

$z_{spec}=(\lambda_{measured}-\lambda_{rest})/\lambda_{rest}$ $z_{photo}=z(C,m)$

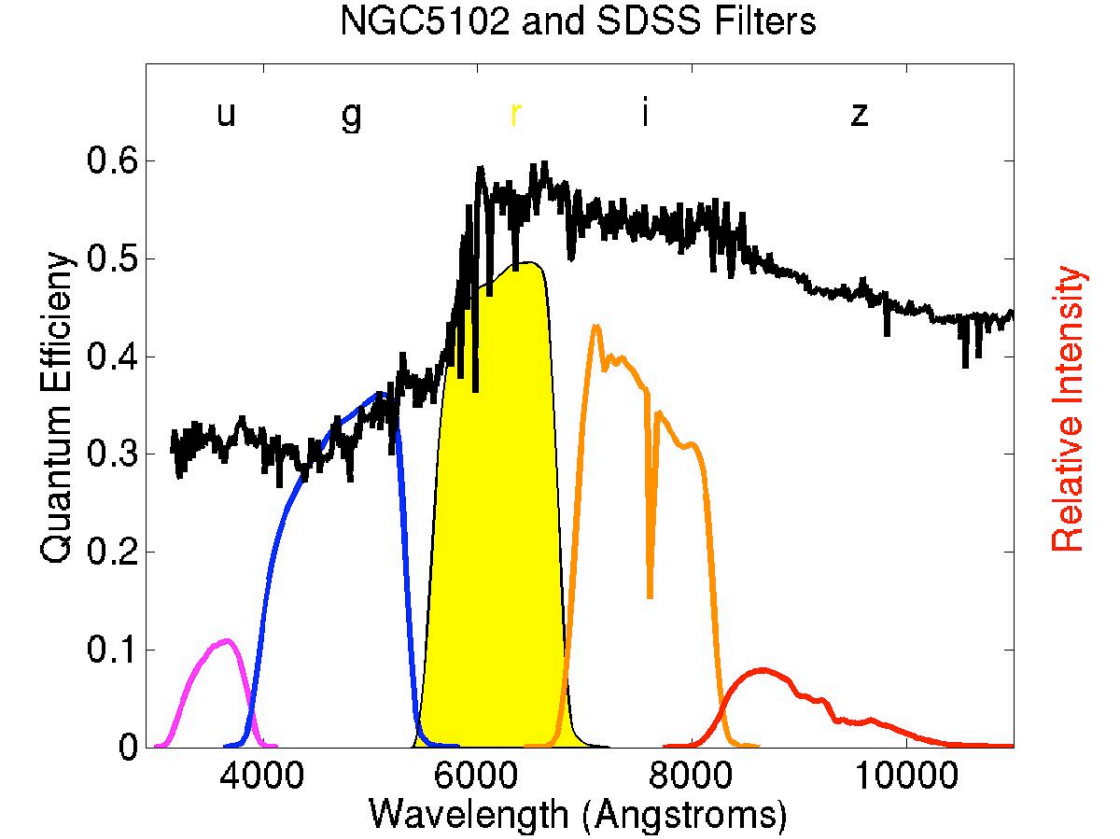
z~0.06 (18000 km/s)



What are Photometric Redshifts?

$z_{spec}=(\lambda_{measured}-\lambda_{rest})/\lambda_{rest}$ $z_{photo}=z(C,m)$

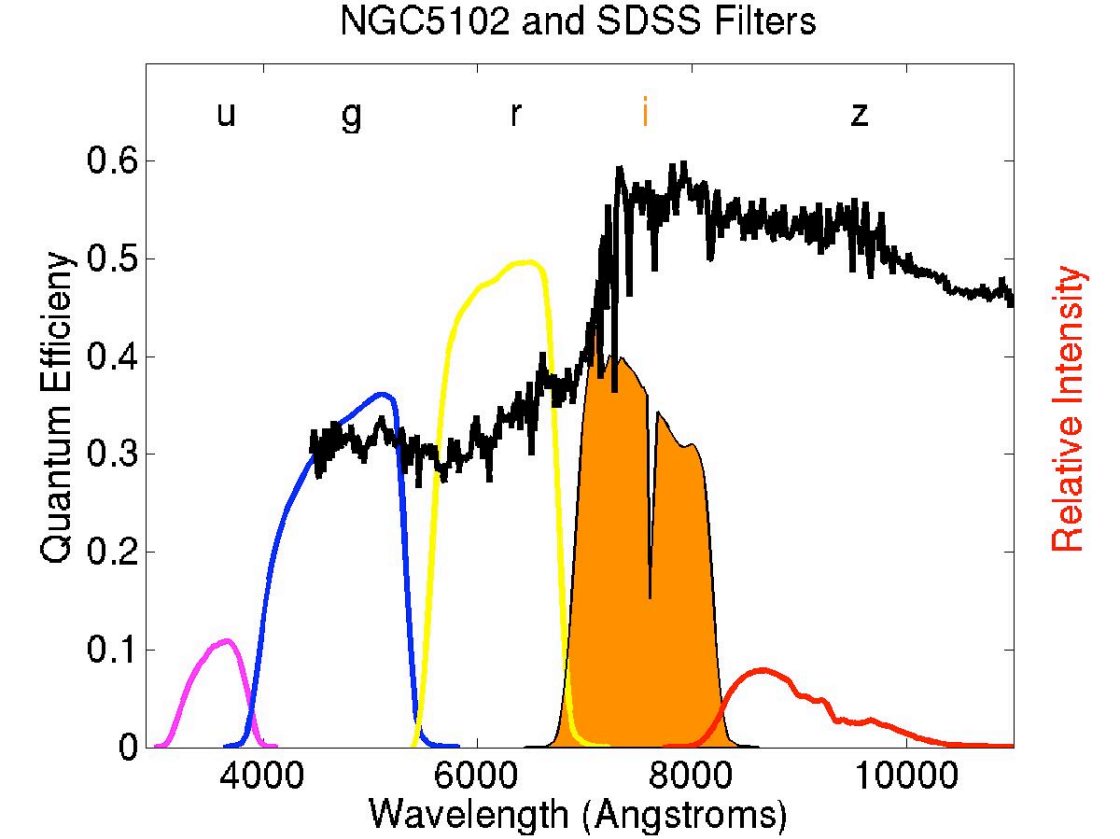
z~0.6



What are Photometric Redshifts?

$z_{spec}=(\lambda_{measured}-\lambda_{rest})/\lambda_{rest}$ $z_{photo}=z(C,m)$

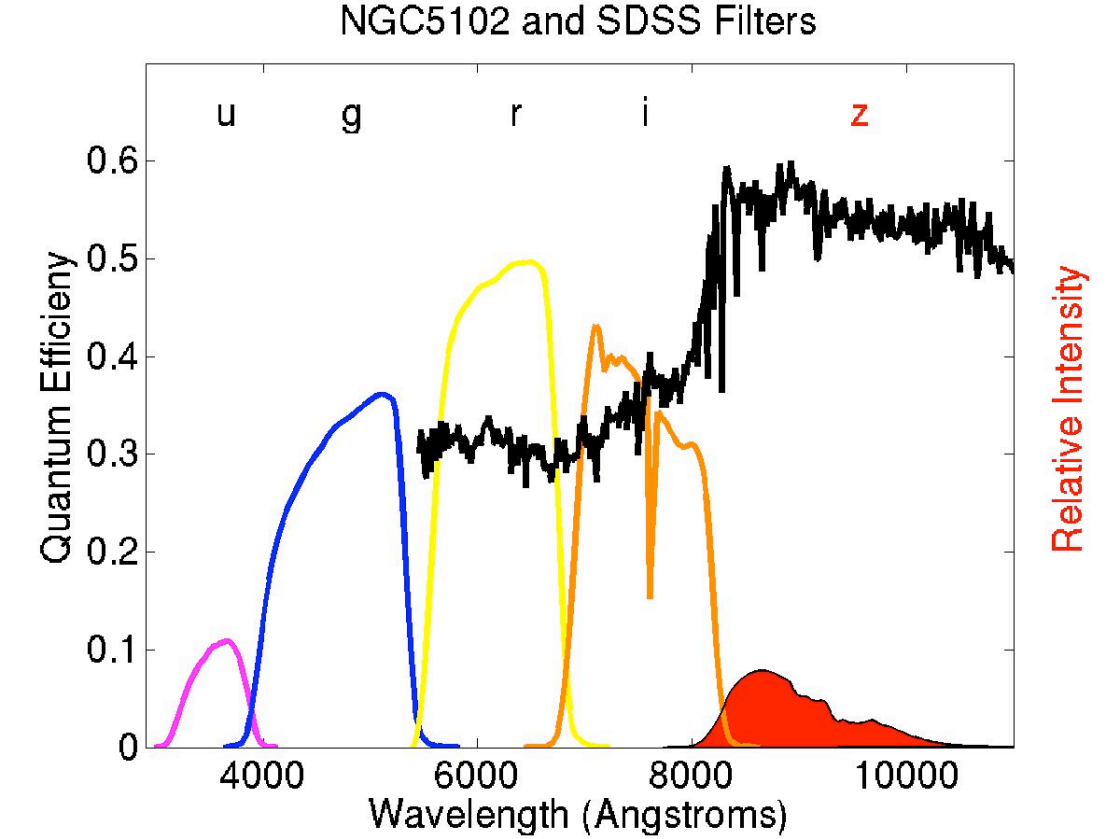
z~0.90



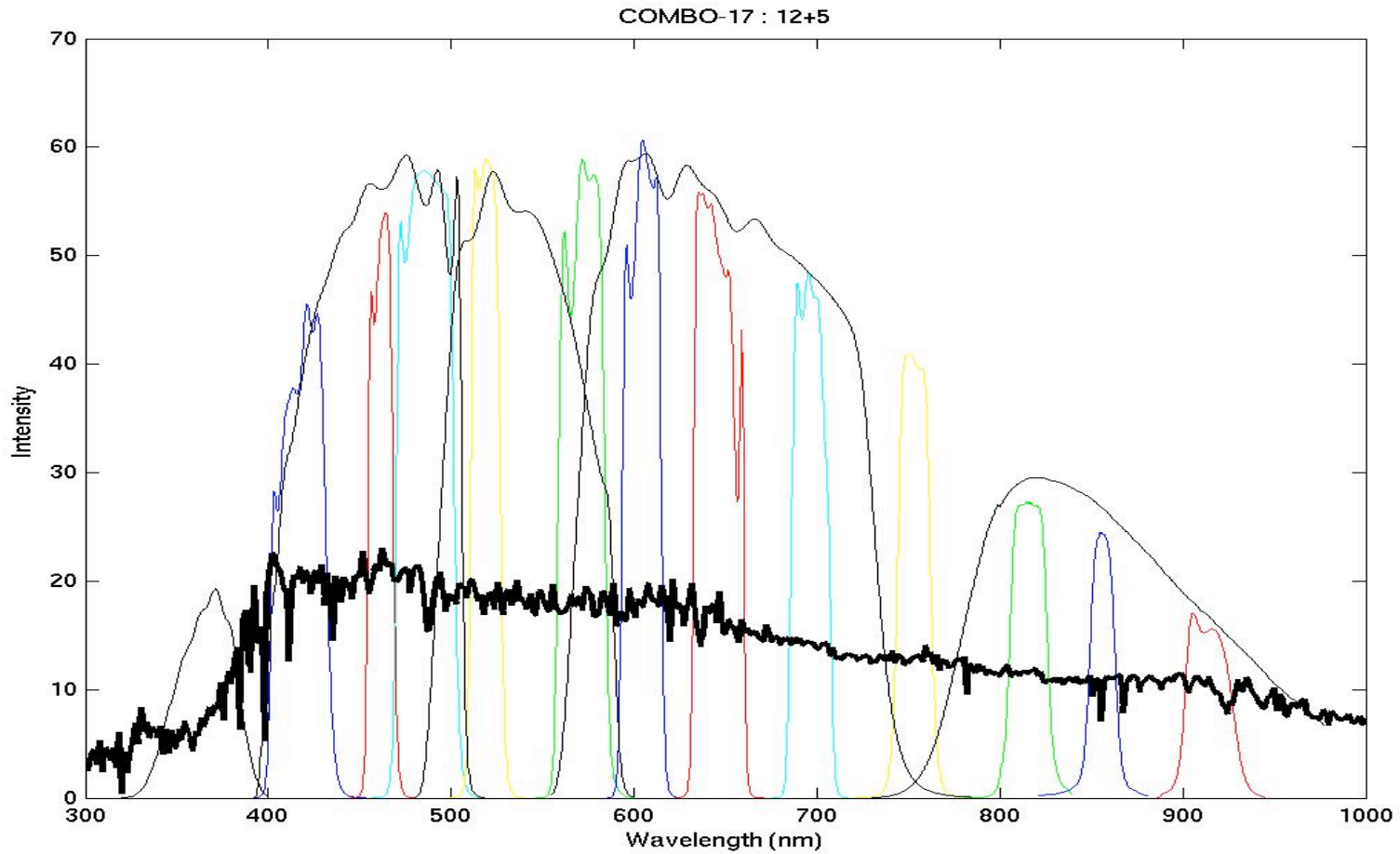
What are Photometric Redshifts?

$z_{spec}=(\lambda_{measured}-\lambda_{rest})/\lambda_{rest}$ $z_{photo}=z(C,m)$

z~1.10



COMBO-17 (ESO 2.2m WFI)



Stretching Photon Use

Photometric **vs** Spectroscopic Redshifts

- BUT: Accuracy is low:
  - **As much as 2 orders of magnitude lower than typical redshift estimates for broad-band photometry**
- YET: There is still science to be done

Photometric Redshift Science?

A couple of applications:

- Cosmology (e.g. Dark Energy Survey, LSST)
  - Weak Lensing/Cosmic Shear (arXiv:0712.1562v1)
  - **Large Scale Structure detection in wide field multi-band imaging surveys (2MASS, SDSS)**
- Deep pencil beam imaging surveys (HDF, HUDF, DEEP2, GROTH Strip, etc)

Photo-z for wide fields

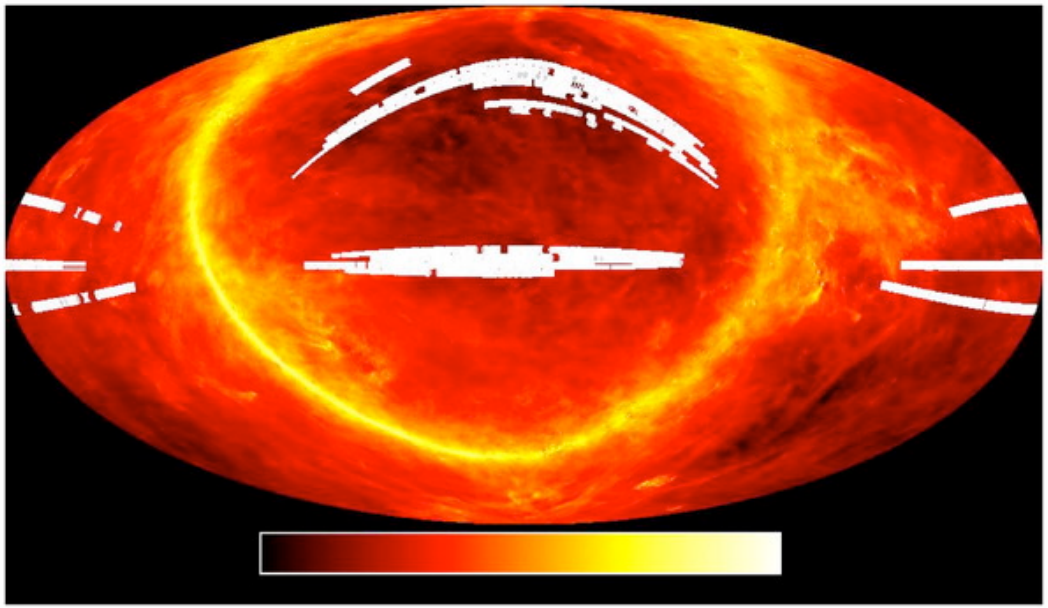
The 3 most prominent wide field surveys used today:

- The Palomar Obs Sky Survey [POSS] (1950-57, 1970s, 1980s)
  - The only full sky optical imaging survey as of today (1m telescope)
  - Was done in two band-passes using glass photographic plates
- The Two Micron All Sky Survey [2MASS] 1997-2001
  - The largest Near IR full sky survey of the sky (1.3 meter telescope)
- The Sloan Digital Sky Survey [SDSS] 2000-now
  - Multi-band CCD imaging of 1/3 of the sky
  - Includes follow-up spectroscopy to shallow depth
- Lets take a closer look at the SDSS and why it is the optimal survey for wide-field Photometric Redshifts today…

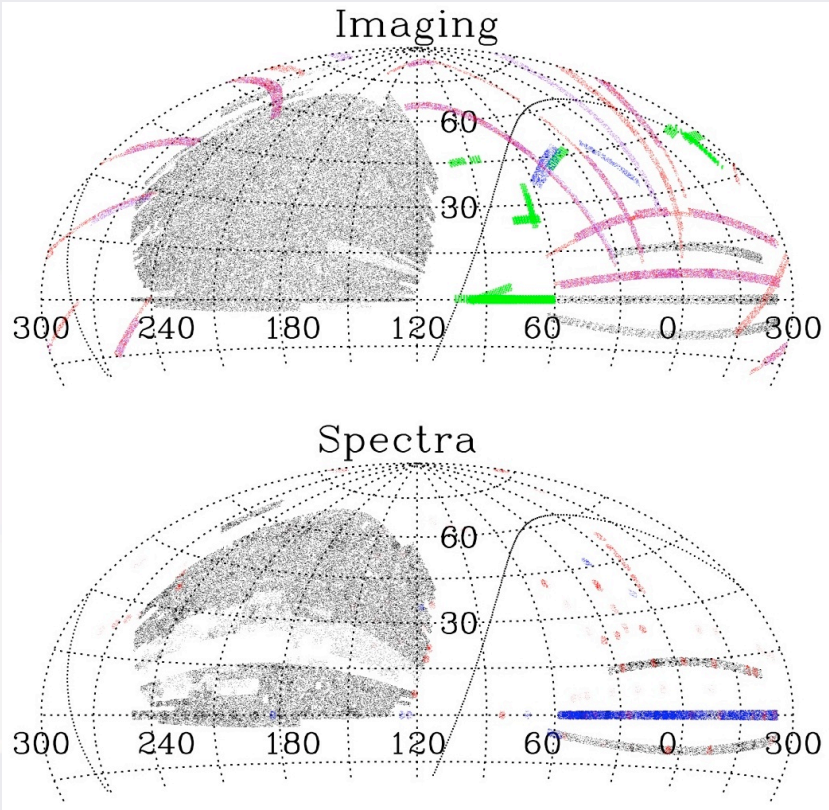
The Sloan Digital Sky Survey

- >9500 deg^2 in 5 bands (u g r i z) to r~22.5, (37GB/hr)
- Images=10TB, MS-sql DB=4TB
- Spectra=1.6x10^6, 8x10^5 galaxies (depth r~18)
- 230GB spectra+data products
- 287 million unique objects

Data Release 1



Data Release 6



Example of a SDSS Query

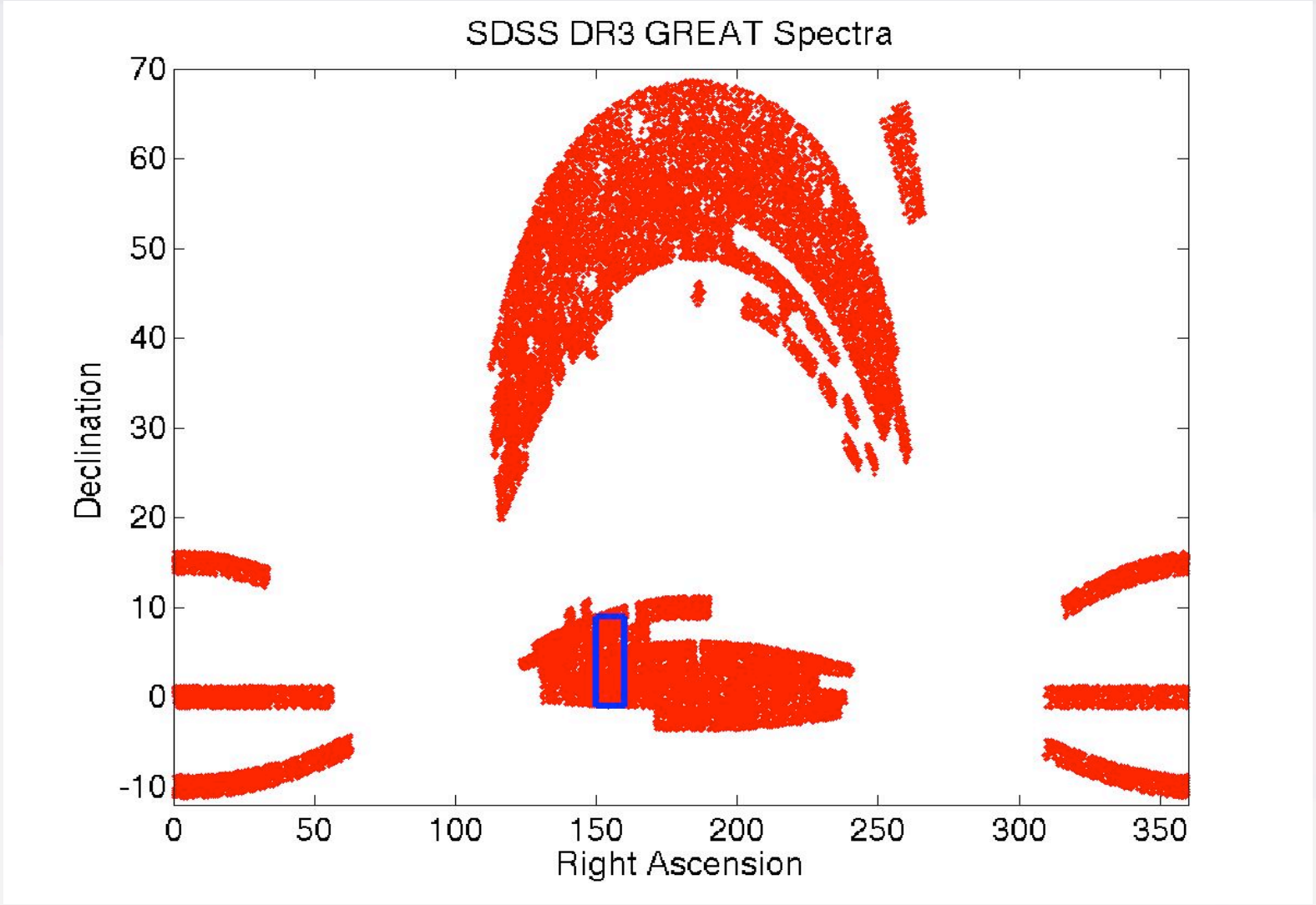
```
Select p.ObjID, p.ra, p.dec
p.dered_u, p.dered_g, p.dered_r, p.dered_i, p.dered_z,
p.petroR50_r, p.petroR90_r, p.fracDev_r, p.q_r,
p.Err_u, p.Err_g, p.Err_r, p.Err_i, p.Err_z,
p.petroR50Err_r, p.petroR90Err_r, p.qErr_r,
s.z, s.zErr, s.zConf
FROM SpecObjAll s, PhotoObjAll p
WHERE s.specobjid=p.specobjid and s.zConf>0.95
and (p.primtarget & 0x00000040 > 0)
and ( ((flags & 0x8)=0) and ((flags & 0x2)=0)
and ((flags & 0x40000)=0) and ((flags & 0x10)=0)
and ((flags & 0x1000)=0) and ((flags & 0x20000)=0) )
```

The Sloan Digital Sky Survey

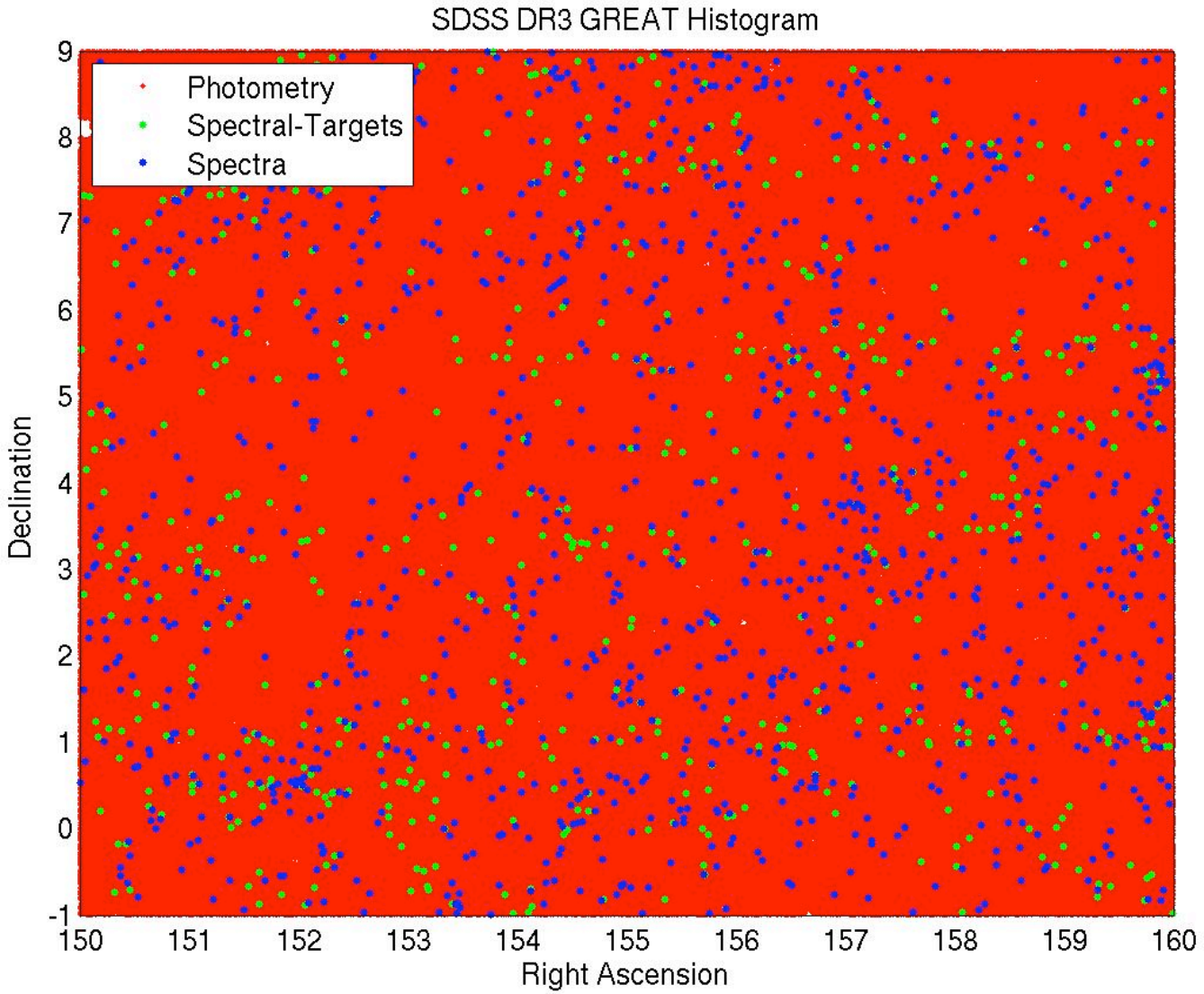
- Obviously one obtains many more galaxies per unit area with photometry versus spectroscopy for a given exposure

- Lets look at the number density of galaxies in the SDSS for photometric versus spectroscopic results

SDSS Data Release 3 (DR3)



SDSS DR3 Photometry+Spectra



SDSS Magnitude Histogram r <=22

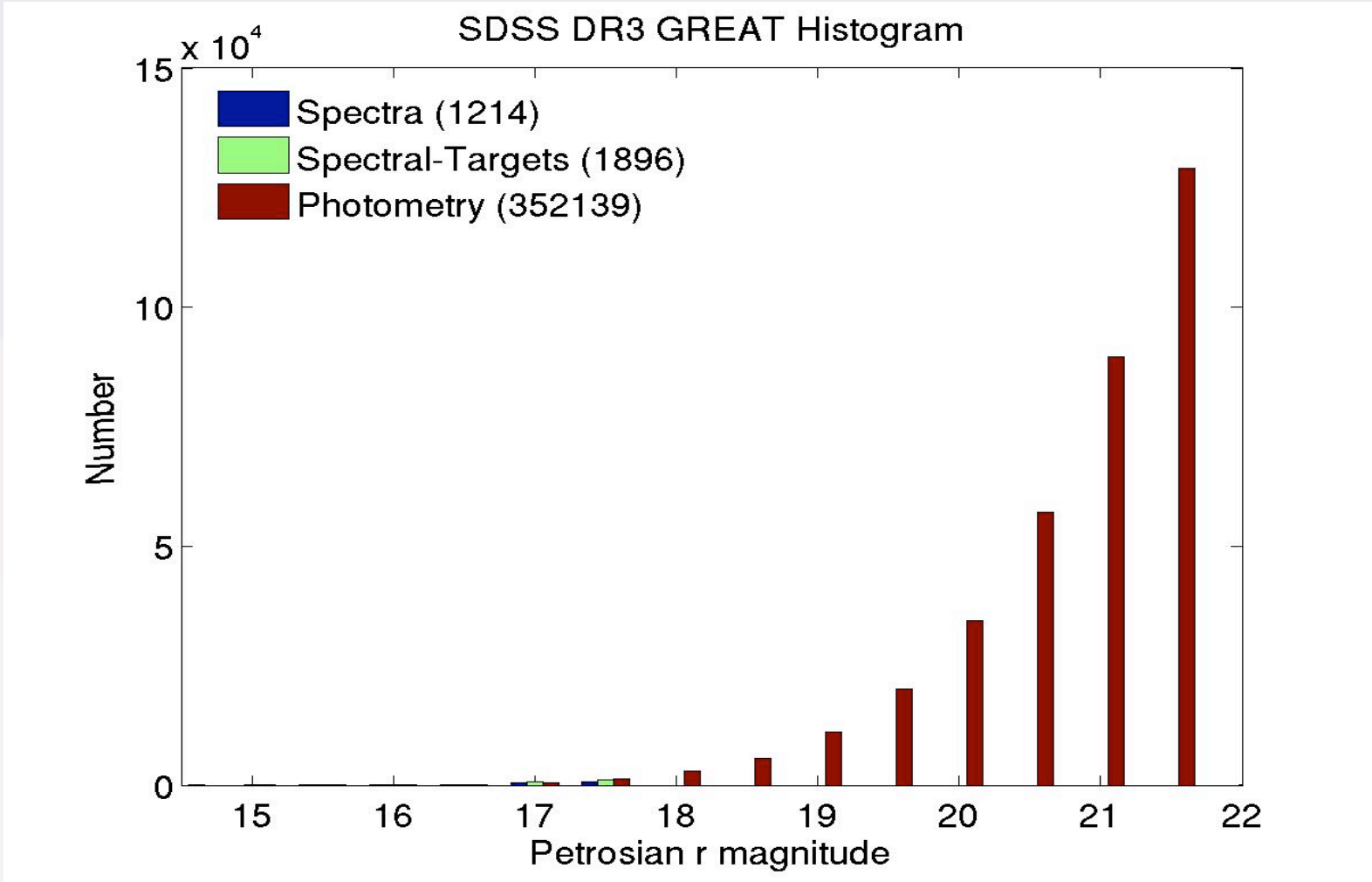


Photo-z methods

Now lets look at some Photometric Redshift methods.

For ~30 years astronomers have used two methods for redshifts on the cheap ➔➔

Photo-z methods

1.) Spectral Energy Distribution (SED) Fitting:

- model based approach
- uses redshifts derived from spectra of artificial galaxies (e.g. Bruzual & Charlot)

2.) Training-Set methods:

- **empirical approach**
- **uses *spectroscopic* redshifts from a sub-sample of galaxies with the same band-pass filters**

Photo-z The Empirical Approach

Training Set Methods need a sub-sample of Galaxies:

- of known spectroscopic redshift
- with a comparable range of **magnitudes** **(**u g r i z**)** to our Photometric survey objects
- These will be our "Training Samples"

"Training Set" Methods

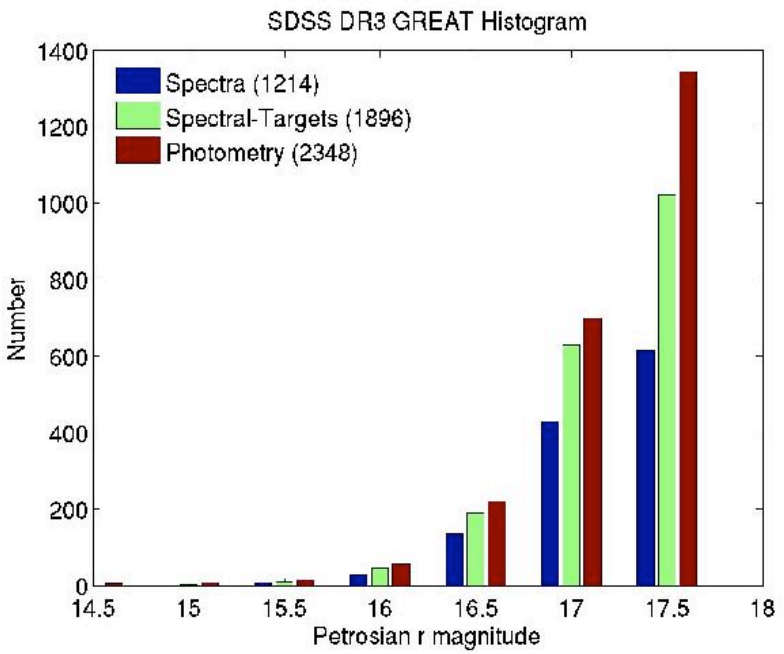
We will need *many* training samples (10,000s), why?

- Not all Galaxies are the same: e.g.
  Spiral, Elliptical, Star Burst, Active Galactic Nuclei …
- They will have many different redshifts

SDSS Image of Hickson 88

Albany 08

Redshifts measured in small SDSS field



"Training Set" Methods

Galaxy Photometric Redshift Prediction History

- Linear Regression was first tried in the 1960s
- Quadratic & Cubic Regression (1970s)
- Polynomial Regression (1980s)
- Neural Networks (1990s)
- Kd Trees & Bayesian Classification Approaches (1990s)
- Support Vector Machines & GP Regression (2000s)

Lets review linear regression quickly before we move on

Linear Regression

The start of Regression: A History in brief!

- Earliest form was the method of least squares
- First described by Gauss in 1794 (he was 18). Used it in 1801 to predict the orbit of the asteroid Ceres
- Gauss **finally** published it in 1809 in his work on celestial mechanics: "Theoria Motus Coporum Coelestium in sectionibus conicis solem ambientium"
- Independently derived by Legendre 1805 & Adrian 1808

Linear Regression

Linear Regression Reminder for our case:

- Models the relationship between a dependent variable y and independent variables X_i, i=1,2,…n

$$y = b0 + \sum_{i=1}^{n} b_n X_n + e$$

- y = galaxy spectroscopic redshifts
- X = 5 broad band pass filter measurements for those galaxies with a measured spectroscopic redshift (y)

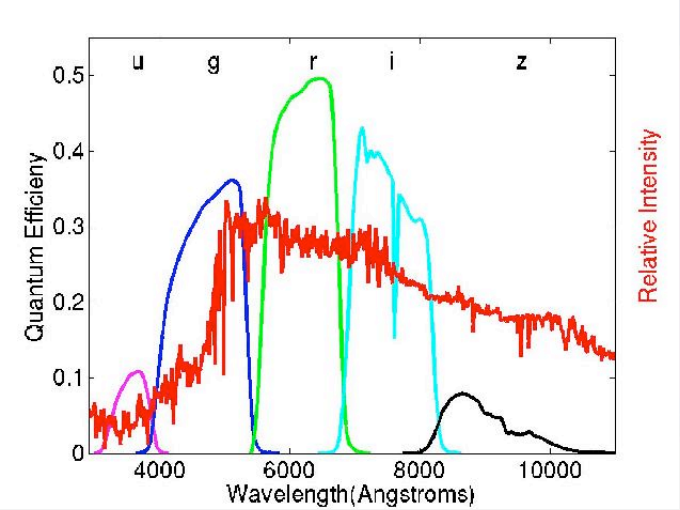
Multiparametric Fitting Methods

Linear regression in the SDSS:

$$\sum_{j=0}^{ngal}(sz_j) = A + Bu_j + Cg_j + Dr_j + Ei_j + Fz_j + e$$

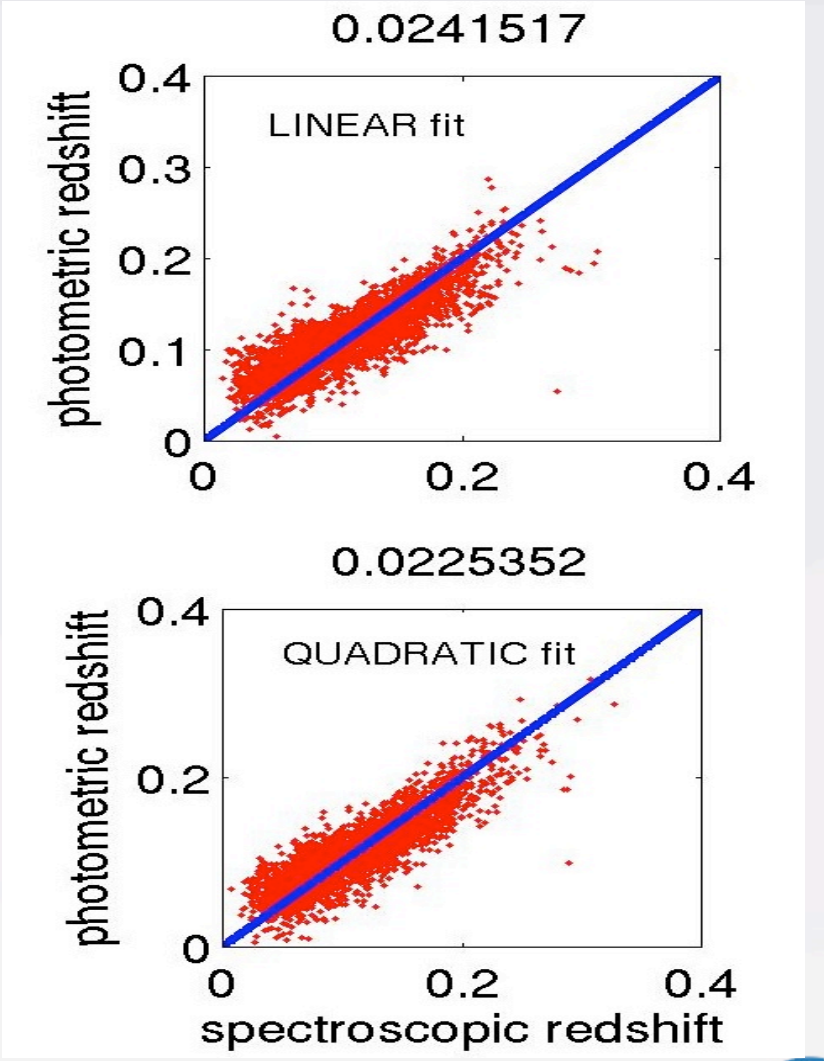
u g r i z = 5 SDSS filters ➔

Solve A,B,C,D,E,F => Photo-z



Linear and Quadratic Fits – Problems?

- Over fitting issues (small N) – hard to quantify.
- No estimates of individual photo-z errors.



Non-Linear Fitting methods

Non-linear fitting methods in use today:

- Quadratic and Cubic Fitting
- Back propagation Neural Networks: (NN) (e.g. ANNz by Collister & Lahav 2004)
- Support Vector Machines (Wadadekar ‘05)
- Bayesian approaches (astro-ph/0607302), etc.

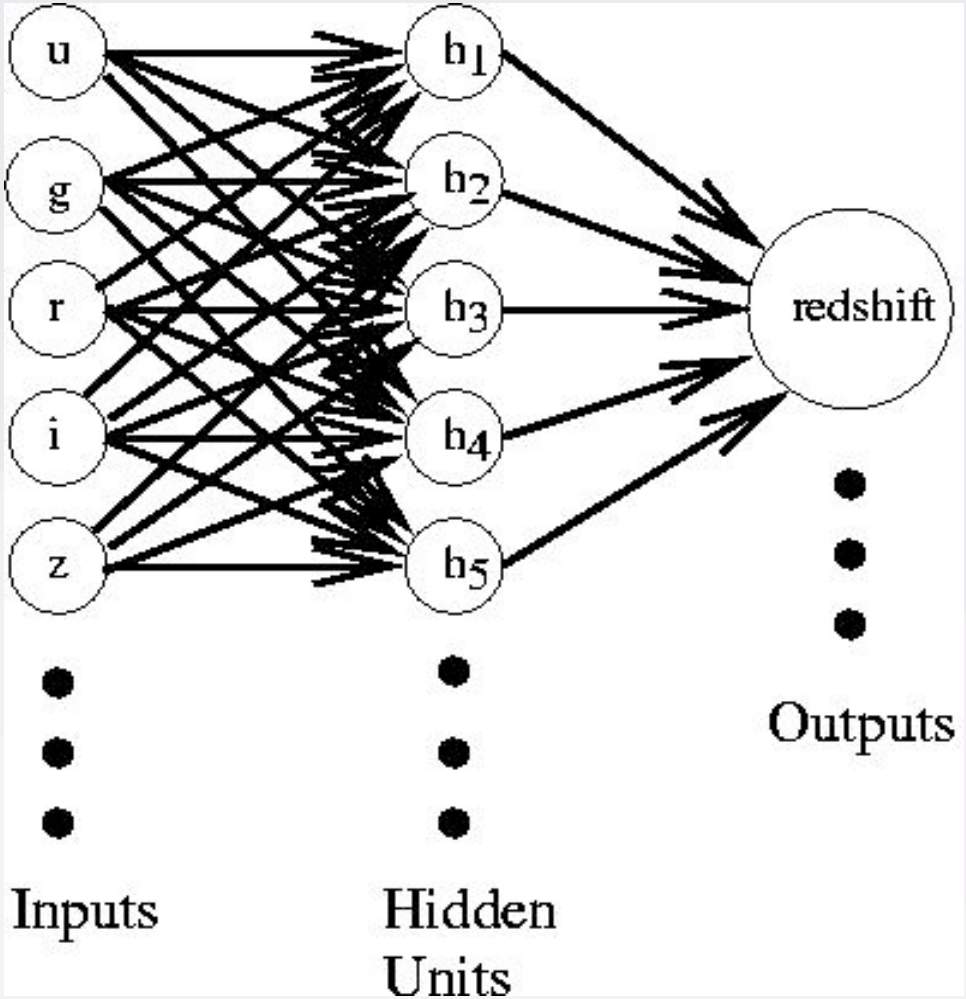
Non-Linear Fitting methods

Our Collaborative Approach to the SDSS

Virtual Sensors:

- Neural Network using Ensemble Modeling (EM)
- Gaussian Process Regression (GP)
  - GPs with reduced rank matrix inversion estimators

Neural Network diagram



Neural Networks

Advantages of NN over simpler methods:

- They **can** avoid over fitting of the data
- It is possible to get error estimates on the predicted redshifts
- They are scalable to large datasets: 10^5-10^6

Disadvantages of NN:

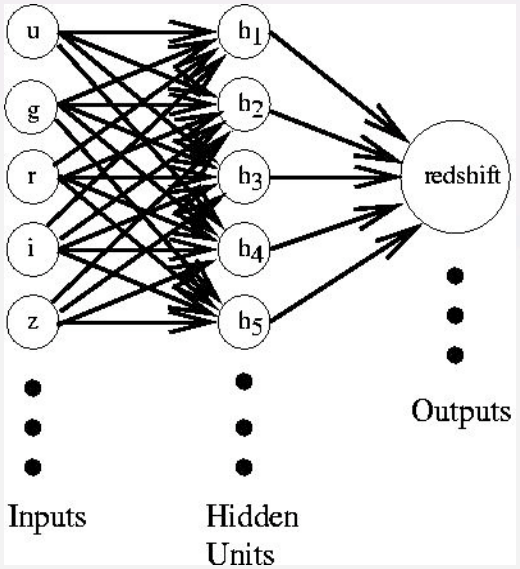
- Often not used properly - rely upon 1 model run. [Use Ensemble Modeling]
- Complaints of their being “Black Boxes”
- Large CPU time requirements when bootstrapping
- Underestimate galaxy photo-z errors

Gaussian Process Regression fitting

Gaussian Process Regression ⇔ Kernel Methods

Kernel Methods have replaced Neural Networks in the Machine Learning literature

WHY?: given a large # of hidden units => GP (Neal 1996).



$h_n > 100$

→→→→→

Gaussian Process Regression

Gaussian Process Regression has a long history:

- Time Series Analysis in Astronomy (1880)
- Military trajectory predictions (1940)
- Geostatistics (1963)

See Mackay (1998) for more information.

Kernel Methods - Gaussian Process Regression

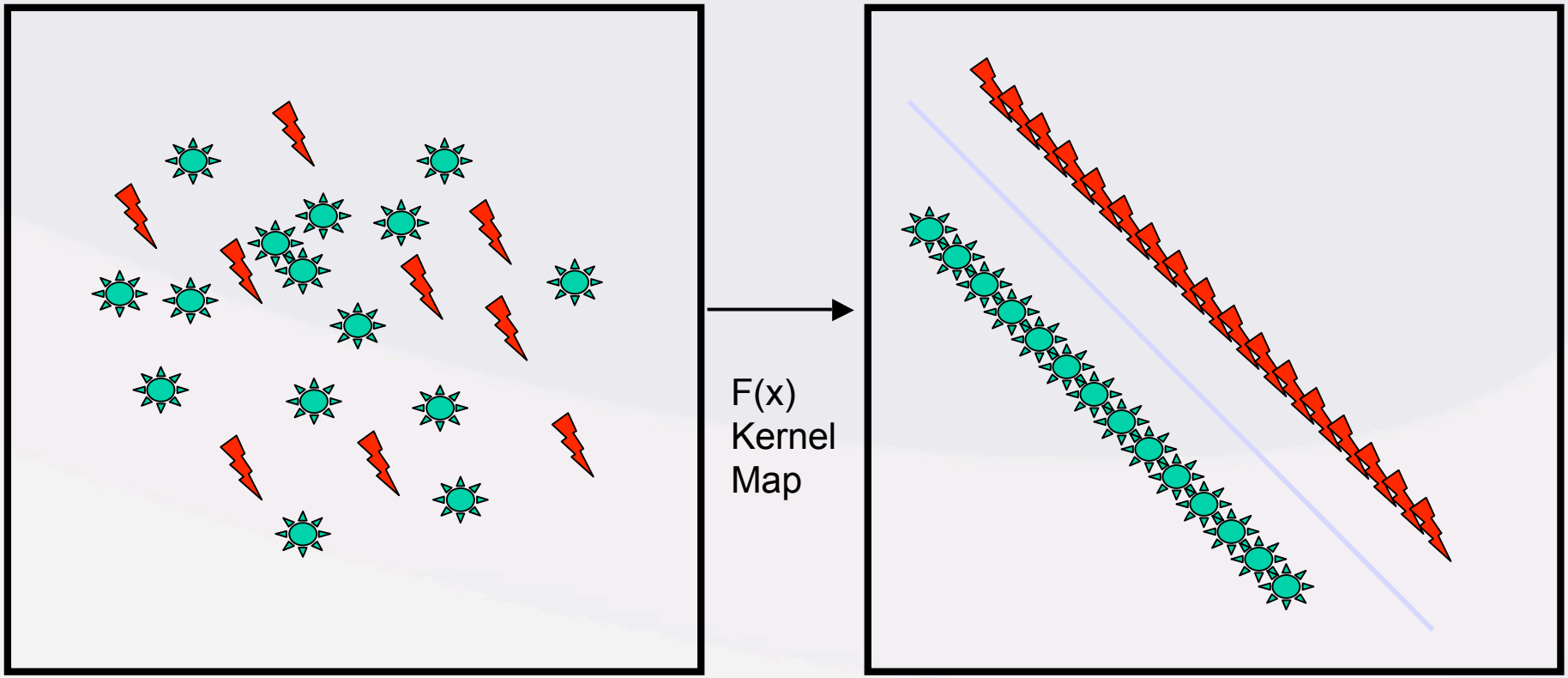
GP regression builds a linear model in a very high dimensional *parameter space* ("feature space" → Hilbert space).

- One can map the data using a function F(x) [kernel] into this high (or infinite) dimensional *parameter space* where one can perform linear operations.

The value of kernels

Original Data without Kernel

Mapped Data using Kernel



Data in original space: highly complex decision boundaries.

Data in high dimensional feature space after mapping through F(x) can yield simple decision boundaries.

GP Regression (Kernels)

GP Advantages:

- Small input data training samples (good for higher redshifts) yet low errors
- Over fitting is eliminated by use of proper priors
- Realistic estimation of individual redshift errors

GP Regression

GP Disadvantages:

- Possibly large CPU time requirements
  - The Kernel (Covariance Matrix) **can** be large:
    $K=(\lambda^2 I+XX^T)^2$ if $X=5\text{x}180{,}000$ (our case) then
    K is a matrix 180,000 x 180,000 and we have:

$$y^* = K^*(\lambda^2 I + K)^{-1} y$$

  - Need to invert this large K matrix - $O(N^3)$ operation
- Kernel Selection is ambiguous (Bayesian like?)
- Black-box like?

GP Regression How-To

Using GPs Part I: Pick a transfer/covariance function

Matern Class Fcn

$$k(r) = \frac{2^{l-v}}{\Gamma(v)}\left(\frac{\sqrt{2vr}}{l}\right)^{v} J_v\left(\frac{\sqrt{2vr}}{l}\right) \qquad v \to \infty$$

Radial Basis Fcn

$$k(r) = \exp\left(\frac{r^2}{2l^2}\right)$$

Rational Quadratic

$$k_{RQ}(r) = 1 + \left(\frac{r^2}{2\alpha l^2}\right)^{-\alpha}$$

Polynomial

$$k(x, x') = \left(\sigma_o^2 + x^T \sum_p x'\right)^p$$

Neural Nets

$$k_{NN}(x, x') = \frac{2}{\pi}\sin^{-1}\left(\frac{2x^T \Sigma x'}{\sqrt{(1 + 2x^T \Sigma x)(1 + 2x'^T \Sigma x')}}\right)$$

GP Regression How-to

Using GPs Part II: That matrix inversion…

With our SDSS (DR3) spectroscopic sample (180,000 galaxies) the matrix size is 180,000 x 180,000

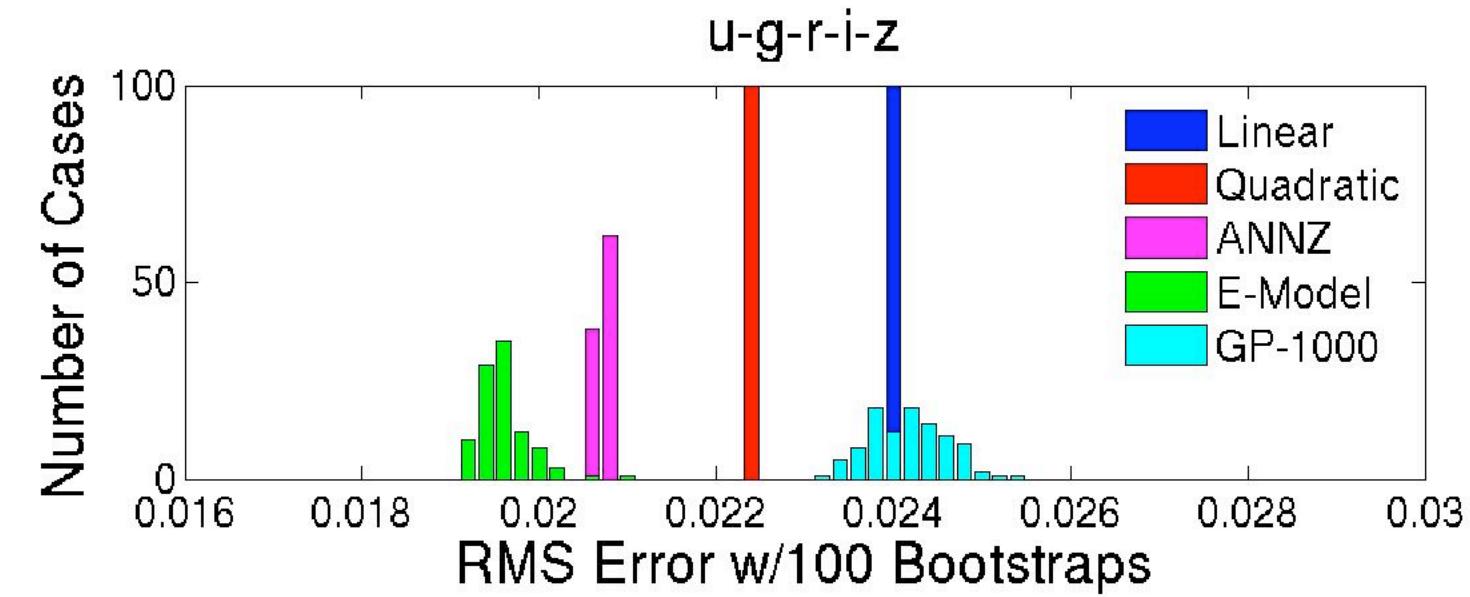
- Need a supercomputer with a LOT of ram and cpu time?
- One can take a random sample of ~1000s galaxies & invert that while bootstrapping n times from full sample (Paper I)
- **However, some *low-rank matrix approximations* work well** (Cholesky Decomposition, Subset of Regressors, Projected Process Approx, etc.)

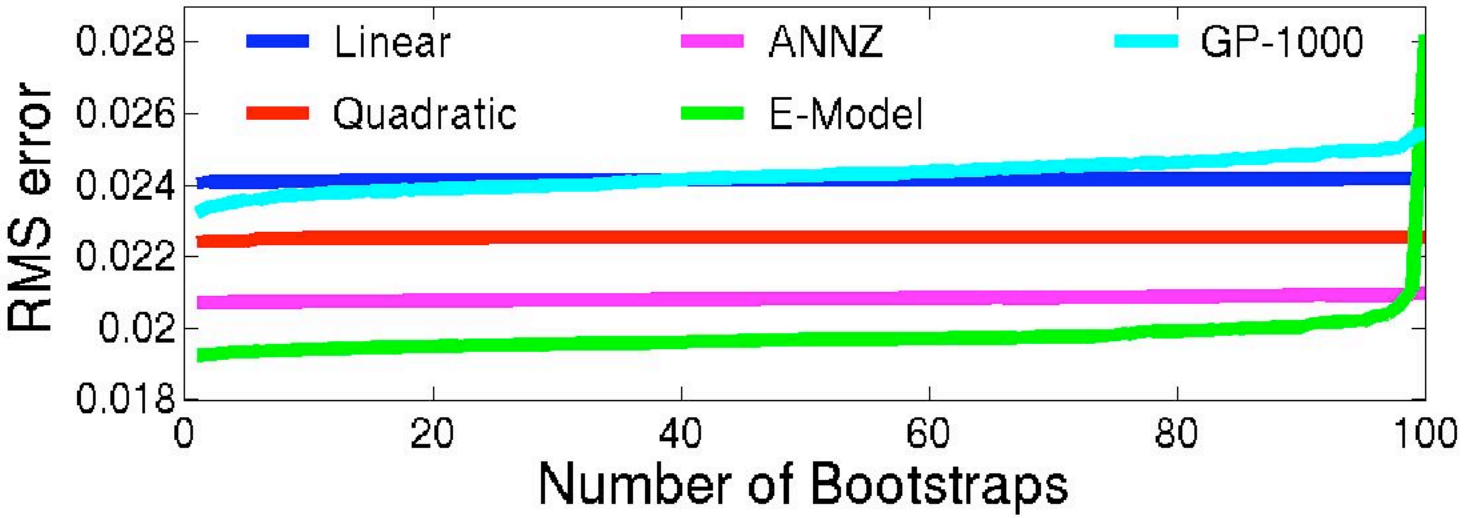
GP Regression (Results)

Results from the SDSS (DR3)

- Compared linear, quadratic, Neural Networks and GPs on the SDSS dataset (only 1000 trainings galaxies for the GPs)
- With 1000 samples GPs performed well especially given their (small) training sample size compared to the other methods
- With *low-rank matrix approximations* GPs performed better than all other methods

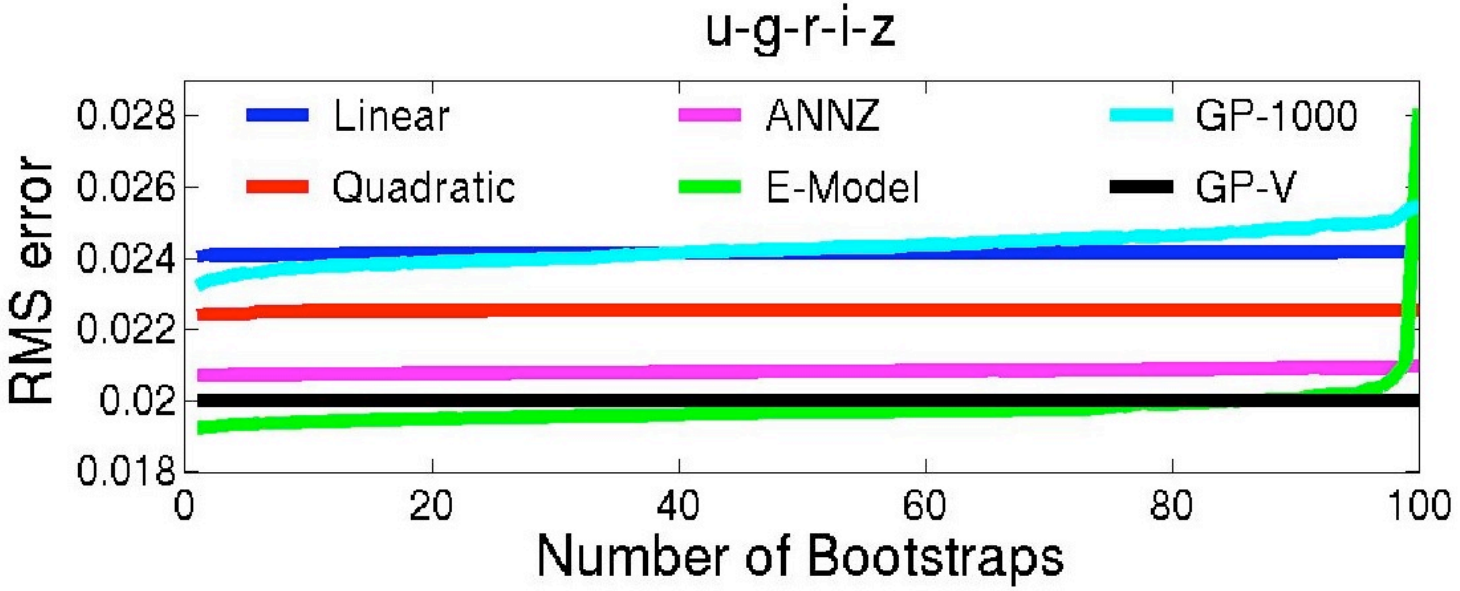
Results: Comparing Methods

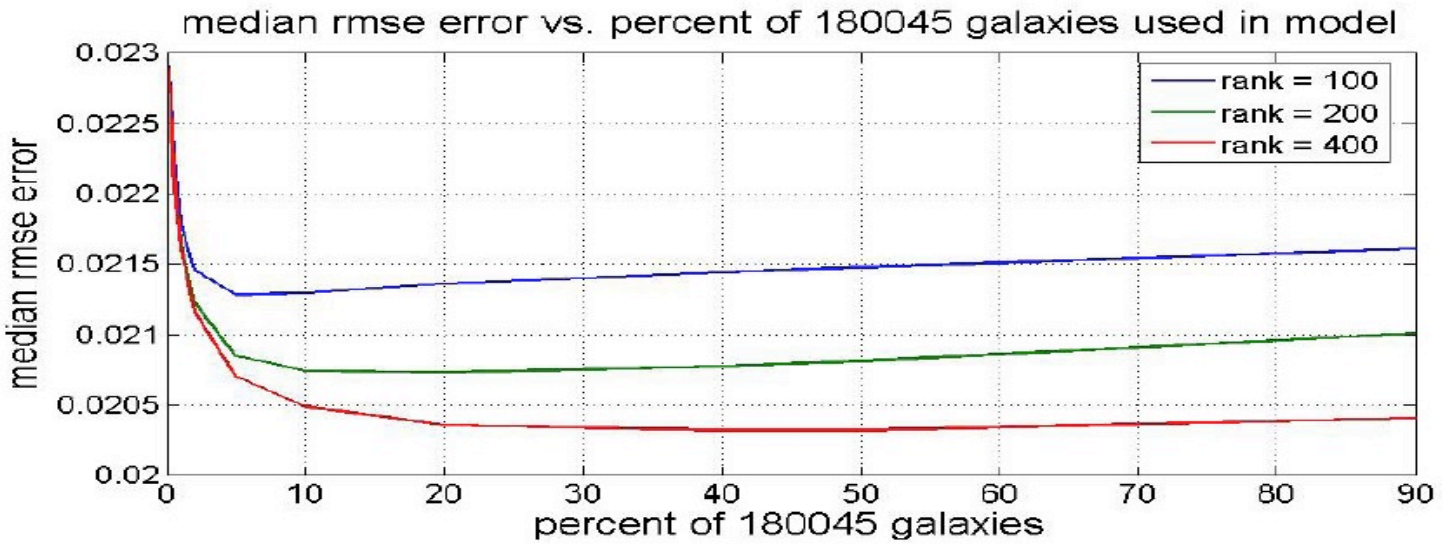




Results: Comparing Methods

➔ GP-V: Rank=1000 for 36000





➔ Beyond 20% (36000) Rank 400 is fairly flat

Results: Other authors

Method Name	σ_{rms}	Dataset[1]	Inputs[2]	Source
CWW	0.0666	SDSS-EDR	ugriz	Csabai et al. (2003)
Bruzual-Charlot	0.0552	SDSS-EDR	ugriz	Csabai et al. (2003)
ClassX	0.0340	SDSS-DR2	ugriz	Suchkov et al. (2005)
Polynomial	0.0318	SDSS-EDR	ugriz	Csabai et al. (2003)
Support Vector Machine	0.0270	SDSS-DR2	ugriz	Wadadekar (2005)
Kd-tree	0.0254	SDSS-EDR	ugriz	Csabai et al. (2003)
Support Vector Machine	0.0230	SDSS-DR2	ugriz+r50+r90	Wadadekar (2005)
Artificial Neural Network	0.0229	SDSS-DR1	ugriz	Collister & Lahav (2004)

Immediate Future Directions

- Use Bruzual-Charlot galaxy population synthesis models to create training-sets for r >18 SDSS photometry
- Use redshifts from DEEP2, VVDS, etc to create training sets for r >18 photometry
- Also use Bruzual-Charlot models for higher-z studies (e.g. Groth Strip, etc)

Conclusions

Astronomy needs good Photometric Redshifts now (SDSS, HDF) and in the future (LSST)

GPs are a competitive way to do regression to get them

GPs avoid over-fitting issues

GPs give robust estimates of individual Photo-z errors

They work well even with small subsamples (high-z)

And …

Conclusions

Astronomers & Comp Scientists can and will continue to work together to solve interesting problems!

Astronomy Data in Context

Astronomy the photon poor field?

Mission/Project	**Data Rate**	**Total Collected**
WMAP (now) [DSN]	0.7Mb/s (16min/d)	30GB/year
2MASS 1m(1998-01)	1Mb/s (~8hr/d)	4TB/year/Telescope
SDSS (Spectra)	12 spectra/min(4000/d)	300GB/year 20000/yr
MRO (Now) [DSN]	0.5-4Mb/s (10hr/d)	800GB/year
MODIS: Terra/EOS	3-10Mb/s (5Mb/s)	19TB/year
SDSS 2.5m(Imaging)	82Mb/s	100TB/year
LSST 8.4m (2014?)	3GB/s	7PT/year
LHC (2008)	40TB/s/inst	5PT/experimt/year